ISSN: 2641-2039
DOI: 10.33552/GJES.2020.05.000611
Global Journal of Engineering Sciences
Iris Publishers

Opinion

Some Thoughts on Reliability of Diagnoses by Human Versus by Machine

Ildikó Ziegler*

Quality Assurance Unit, Vanessa Research Ltd., Hungary

***Corresponding author:** Ildikó Ziegler, Quality Assurance Unit, Vanessa Research Ltd., Hungary.

Received Date: May 07, 2020

Published Date: May 14, 2020

Opinion

Not long ago the Comment by Tessa S Cook [1] appeared in Digital Health following the informative article Xiaoxuan Liu & colleagues [2]. The question of whether deep learning is more reliable/ accurate than human health-care professionals in detecting diseases from medical imaging was in the focus of the papers.

The Comment [1] added much to the discussion since the author listed several factors those influence the accuracy of diagnosis made by A.I. or a human.

The actual discussion speaks about diagnoses based on medical imaging techniques [1,2]. I believe the consideration of these factors can be continued by estimating the tendencies of these factors in the long run and actually, it may bring some inspiring thoughts, not only in connection with medical imaging, but in general, regarding the learning capabilities of artificial intelligence.

Theoretically the following error types could be identified (see e.g. [3,4]):

- Deviation occurring due to the lack of expert consensus.
- False positive diagnoses
- False negative diagnoses
- Other accidental errors
- Other systematic errors

We may make an estimation about the tendency of these errors in time one by one based on the statistical nature of these types of errors (see e.g. [5,6]). Supposing that the capacity of A.I. to handle data and to conduct repetitive calculations/ decisions is always kept greater than the capacity requirements of actual tasks, that means no technical limitation of processing the growing data set occur in time during the development of A.I.

Moreover, a single person, theoretically, cannot grow her/ his knowledge in an unlimited manner, partly, because we, humans, forget things, partly, because other circumstances, e.g. sickness or death may prevent us from continuing our learning process. However, the human being as a society also can be considered as a continuous learner group and as such the achievable total knowledge of human kind seems to be unlimited.

Based on the assumptions above, let's take a look at the mentioned influencing factors one by one:

Systematic errors, others the mentioned in the earlier groups are the characteristics of the given A.I., and the way it functions. It cannot change during the learning process, however it may step wise decrease if it is detected and more or less handled, for instance by upgrading the software, developing the algorithm, etc. In case of a given A.I. it is constant during the learning process, but it will decrease during the development of A.I. technology over time.

Accidental errors (others than the mentioned in the earlier groups) are characterized by Gaussian distribution [3-6]. It means that with the elapsed time the standard deviation characterizing of the learning process of a given A.I. will decrease during the learning period.

False negative and false positive results mean that the decision about the diagnosis is not true.

In those cases when results (diagnoses) are not independent from each other - and e.g. biostatistics or clinical decisions are this

kind – the deviations are characterized by Bayesian distribution [7,8]. The situation with this type of error is similar to the previous one: as the number of items in the data set on which the learning is based increases, the probability of the error decreases. With the elapsed time the A.I. is still expected to improve its effectiveness in the long run.

The most complex problem is to estimate the tendency of those diagnoses which medical doctors and specialist have different opinions without consensus. It has to be taken into account that a human solves problems in a different way compared to A.I. Deep learning is based on a super computer programmed with sophisticated algorithms, thus, it always uses linear deductions based on large dataset and mathematical logic, while we, humans tend to use our creative, associative thinking as well as checking the associative ideas based on professional rational. It is mainly expressed – as T. Cook [1] also pointed out – by striving for a holistic approach: "medical practitioners begin with the history of the present illness, the review of systems".

The use of creative thinking may shorten the time requirement of problem solving and also make serendipitous results possible. As long most studies take the approach of evaluating diagnostic accuracy in isolation without reflecting to clinical practice, development of the accuracy cannot be expected. Conversely, if diagnoses are combined with subsequent experience in the underlying data set, and this data set contains a sufficiently large number of cases, deep learning may provide a better approach to issues that have previously aroused professional debate.

At this point, A.I. may start to make more accurate diagnoses than human health-care professionals. This approach may be true until some rare disease appears for which there is little experience. In this case, however, the human professionals are also quite likely to misdiagnose, without owning the basis for comparison. It seems the theoretical approach led to the opposite conclusion than Liu et al [2], but it is not the case.

The reason lies in the fact that our theoretical estimate was made long-term, assuming a large amount of initial data and ample storage capacity. However, the size of the currently available dataset is not close to the desired size, for instance, in the present case [1] it meant 25 studies. Consequently, nor did we achieve the reliably large amount of data from which artificial intelligence could statistically infer uncertainty in case of lack of expert consensus.

At the present state of the art, deep learning functions as a kind of secondary measurement method that is calibrated to human experience, as "primary method". As such, a secondary "measurement method" cannot be more accurate than the primary "measurement method" for which it was calibrated.

The commentator [1] also mentions that deep learning acts as a black box: one cannot see the algorithm's decision-making mechanism, "it still cannot tell us why the end result is produced". It is suspected that the everyday practice of the medical profession and the science of programming are so far apart from each other that the intensive team work of good communication, as usual with medical devices, will result in sufficiently reliable futuristic technologies.

Just as the driver does not see how the fuel burns in the cylinder head under the piston, but he believes in the service of car mechanic when his car is deemed suitable, the medical professional will not have a chance to look into the details of the algorithm. It is necessary to trust the multi disciplinary development team who validates the machine learning system they have developed.

It is also interesting to consider Moore's and Kynder's laws. Moor's law [9,10] is an experimental law stating hat the number of transistors on a chip would double every two years / 18 months. According to Kynder's law [10,11] the density / capability of hard drive storage media would double every 18 months. It will be interesting to see in the long run if the size of the data set used for machine learning – analogously to the afore mentioned laws – will follow an exponential curve as a result of technological advances. I am voting in favor.

Acknowledgment

On leave from Total Quality Mangement, Gedeon Richter Plc.

Conflict of Interest

No conflict of interest.

References

1. Cook TS (2019) Human versus machine in medicine: can scientific literature answer the question? The Lancet Digital Health.
2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, et al. (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health.
3. JCGM 100 (2008) Evaluation of measurement data—guide to the expression of uncertainty in measurement.
4. EA-4/16 (2003) EA guidelines on the expression of uncertainty in quantitative testing. European co-operation for Accreditation.
5. Shahbazikhah P, Kalivas JH (2013) A consensus modeling approach to update a spectroscopic calibration. Chemo metrics and Intelligent Laboratory Systems 120(1): 142-153.
6. Wiora J (2016) Problems and risks occurred during uncertainty evaluation of a quantity calculated from correlated parameters: a case study of pH measurement. Accreditation and Quality Assurance 21(1): 33-39.
7. Daniel WW, Cross CL (2018) Biostatistics: A Foundation for Analysis in the Health Sciences, 11th Edition, Wiley.
8. Hoffman JIE (2015) Biostatistics for Medical and Biomedical Practitioners, Academic Press – Elsevier.
9. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8).
10. Walter C Kryder's Law (2005) Scientific American 293(2): 20-21.
11. Esener SC, Kryder MH (1999) The Future of Data Storage Technologies (report). International Technology Research Institute pp. 85.